THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 2, February 1983
Printed in U.S.A.

Growth, Complexity, and Performance of Telephone Connecting Networks

By V. E. BENEŠ

(Manuscript received December 10, 1981)

In an effort to free telephone traffic theory from some of its dependence on independence assumptions, and to reap some benefit from its traditional state equations, a systematic search is made to find relationships between load, loss, size, structure, and other network parameters that are simple, universal, and informative. Three principal topics are covered:

(i) A load-loss-size formula, linking some half-dozen network parameters by a rational function, and used repeatedly to give

(ii) Lower bounds on the number X of crosspoints in networks

(iii) Asymptotic results about blocking, growth, and complexity of selected network structures in passing from finite to "infinite" sources at constant load.

The major results in (ii) imply that for all practical networks on N terminals, the crosspoint count X must grow like N log N, i.e., incurring loss by restricting access or concentrating cannot avoid the N log N growth rate known to be exacted by nonblocking networks. The chief result under (iii) is that as a constant load is spread over N terminals, then the number X of crosspoints needed to keep loss less than $\epsilon > 0$ need grow only linearly with N, at a rate dependent on ϵ, while the usage (erlangs carried per terminal) goes to zero.

I. INTRODUCTION

The relationships between traffic carried and traffic lost, between load and loss, have always been at the center of interest in telephone traffic theory. Since the time of Erlang,[1] over fifty years ago, the principal problems of traffic theory have been analytical: to predict mathematically, from the structure and mode of use of a switching or connecting network, and from the assumed stochastic behavior of the customers, how much traffic the network will *carry* on the average,

and how much it will *lose* as a result of blocking, overload, suboptimal routing, or incomplete searches for paths. As telephone networks have become larger, two more design parameters of interest have emerged and now command attention: the *size* of the network as measured by the number of crosspoints, and its c*omplexity* as measured, for example, by the number of stages of switching it has.

The probabilistic principle that it is very unlikely for more than a moderate number of customers to want to talk simultaneously has been the theoretical basis of traffic theory since its start. We can view it as an unrefined analog of the principle in information theory that separates a relatively small class of events that exhaust most of the probability from a remaining large class of very unlikely events. This principle has led quite naturally to the use of concentrators, and of networks in which blocking, mismatch, and overflow all can and do occur as it were by design. It is a function of traffic theory to articulate this principle in mathematical models for operating telephone networks, and to use such models to examine its implications for the growth and complexity of networks, as well as their loads and losses.

Even for the simplest stochastic models, progress with these tasks and problems has been very slow because of the combinatorial complexity of the network, the very large number of network states, and the lack of approximate methods. Thus, it is particularly important to find relationships between load, loss, size, and other network parameters that are simple, universal, and useful, even for very large networks. They should be simple in, for example, *not* requiring solution of very high-order systems of equations, universal in being relatively independent of network structure, and useful in providing inequalities, estimates of performance, and information about the growth of cost and complexity with network size.

In this paper we try systematically to sketch out some of these relationships and associated ideas. The results are of necessity spotty, and no claim is made of completeness or originality, only of rigor. Three principal topics are taken up here: (*i*) a load-loss formula, linking some half-dozen network and performance parameters by a rational function; (*ii*) lower bounds on the number of crosspoints in a network; (*iii*) asymptotic results about blocking, growth, and complexity of selected network structures in the limit of passage from finite sources to Poisson arrivals, with total offered traffic held constant.

II. SUMMARY

The organization of the sequel is as follows: By way of some background, we start with discussions of blocking, loss, concentration, etc., and of their relation to the basic principles of telephone traffic theory and engineering. After various preliminary sections on model-

ling, we call attention to a (known) generalized Erlang formula that connects some of the important parameters of an operating network. We note its technical consequences, and use them repeatedly in the rest of the paper.

Next we take up the problem of the growth of the number of crosspoints with the number N of terminals. The question we try to answer is this: When can the $N \log N$ order of growth necessary for nonblocking networks be reduced by allowing a fixed, small probability of blocking, using, for example, concentrators, or other forms of incomplete access? The answer is that it cannot, unless we consider a familiar, special kind of low traffic limit in which line usage vanishes. It is shown, quite generally, that networks arranged in stages *must* grow like $N \log N$ if certain (very reasonable and mild) traffic, access, and symmetry conditions are met. This result, similar to known results for nonblocking networks, implies that neither judicious concentration nor a nonzero loss can lower the order of growth from the $N \log N$ exacted by the nonblocking case.

The final sections describe various large networks with a simple structure vis-à-vis blocking; their loss probability can be calculated exactly in spite of the astronomical number of states. These networks are based on such structures as trunk groups, frames, and remote concentrators, all familiar to the traffic engineers. We are interested in studying these exact solutions as we let the network grow while keeping the total traffic constant; this kind of growth amounts to adding more and more customers, each of whom contributes less and less traffic, and results in a passage from finite to infinite sources (a Poisson process of arrivals) at constant offered load. The blocking formulas can be studied in this limit, and they lead to close connections with the classical Erlang function, $E(c, a)$. As an application, we can give methods of synthesizing very large networks with prescribed blocking probabilities. In particular, as a constant offered load is spread over more and more customers, the number of crosspoints sufficient to achieve less than ϵ in blocking need grow only linearly with the number of customers.

There is a bibliography of background reading and related work following the references.

III. STATEMENT OF RESULTS

3.1 Generalized Erlang formula

Using some standard "dynamic" assumptions[2] to describe random traffic, we show that half-a-dozen parameters, all characteristic of network size and performance, are related by a simple, rational function. These parameters are:

N = number of terminals on a side of a two-sided network
= number of inlets = number of outlets
m = (mean) carried load = equilibrium average number of calls in progress
λ = calling rate per pair of idle customers
bl = probability of blocking, from the "wire-chief's" point of view
σ^2 = variance of number of calls in progress,

and the formula states that,[2] very simply,

$$1 - bl = \frac{1}{\lambda} \frac{m}{(N-m)^2 + \sigma^2}. \tag{1}$$

For some purposes the parameters

$p = m/N$ = line usage = erlangs carried per inlet (outlet)
$a = \lambda N^2$ = total offered load (wben everyone is idle)

are more significant or convenient, and using them the formula is recast as

$$a(1 - bl) = \frac{m}{(1-p)^2 + \sigma^2 N^{-2}}.$$

Indeed, there are many ways of twisting and inverting the basic formula, each one illuminating some special aspect; several such will appear later. It can be sbown that if $N \to \infty$ while $a = \lambda N^2$ remains constant, then p and $\sigma^2 N^{-2}$ go to zero, and we have Erlang's original result, the "tautology"

$$a(1 - bl) = m$$

or

$$\text{blocking} = \frac{\text{load lost}}{\text{load offered}}.$$

The formula (1), really a generalization of Erlang's loss formula, is useful for the following applications:

(*i*) Order of magnitude estimates

(*ii*) Asymptotic analyses for large networks, with or without a passage from finite to infinite sources

(*iii*) Bounds on the number of switches in a network

(*iv*) Growth and complexity bounds for networks with given load and loss constraints

(*v*) Synthesis of networks having prescribed parameters.

All these applications are illustrated in the text that follows.

3.2 Concentration

The principle that high occupancy states are very unlikely has

suggested the use of concentration, and it is pertinent to assess the effect and value of concentration in large networks. Some results of this kind are in tbe next three subsections, and they warrant these conclusions:

(*i*) Practical networks must grow like constant $N \log N$, whether there is concentration or not. Concentration affects primarily the value of the constant, and of course the blocking and the carried load.

(*ii*) The extent of possible concentration is limited by the loss and the carried load. As migbt be expected, higher line usage and lower blocking imply less concentration.

3.3 Growth without concentration

In tbe prototypical networks without concentration (e.g., those made of stages of square switches), the number of crosspoints for N customers must grow like $N \log N$ no matter what load is carried. Several arguments are given for similar lower bounds, some purely combinatorial, others involving traffic concepts and parameters. In particular, if blocking is to be kept less than ϵ, and total traffic $a = \lambda N^2$ is kept constant while N increases, the requisite networks must grow like $N \log N$ if they are made of stages of square switches: especially, for s stages

$$X = \text{number of crosspoints} \geq N \log N + sN + \log(1 - \epsilon) - a.$$

3.4 Growth with full access, allowing concentration

Some simple and mild combinatorial properties, possessed by all practical networks, mandate an $N \log N$ order of growth in the number X of crosspoints, even when concentration is permitted. A network provides "full access" if every inlet can reach every outlet by some path. A network is said to be "arranged in stages" (or "made" of stages) if its terminals are partitioned into sets $T_1, T_2, \cdots, T_{s+1}$, such that T_1 consists of the inlets, T_2 to T_s are sets of internal nodes or junctors, and T_{s+1} are the outlets, with crosspoints placed only between T_i and T_{i+1}, $i = 1, \cdots, s$. (See Fig. 1). Here s is the number of stages, and every call traverses each T_i exactly once, in the specified order or its reverse. Finally, a network arranged in stages is called "symmetric" if it looks the same from each terminal in any given T_i; we content ourselves with this informal definition here; a precise one can be given in terms of group theory.[3]

We prove this fundamental telephonic fact: A symmetric network that provides full access and is arranged in stages must have at least

$$en \log N \quad e = 2.71828 \ldots$$

crosspoints, where N is the number of inlets (outlets, too), and n is the "neck size," defined as the size of the smallest T_i:

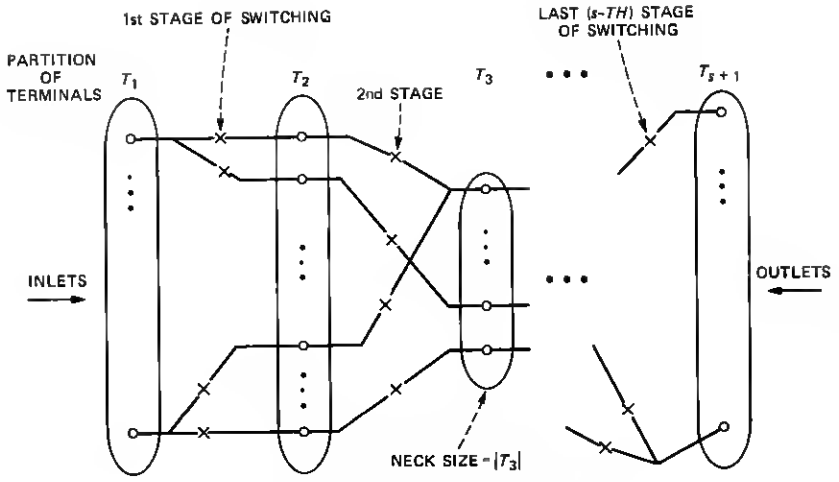

Fig. 1—Network arranged in stages.

$$n = \min_{1 \le i \le s+1} |T_i|, \quad |A| = \text{cardinality of } A.$$

The ratio n/N is a global measure of concentration or expansion. If, in particular, all the T_i are equinumerous, as happens if the network is made of stages of square switches, tben the neck size is N, and there must be at least $eN \log N$ crosspoints, regardless of the traffic characteristics.

3.5 Growth without full access, allowing concentration

The condition of full access is so reasonable that few engineers would consider a network that lacks it. Nevertheless, probabilistic arguments yield $N \log N$ lower bounds for the crosspoint count X even when this condition is dropped. The point is that if the network is to carry a reasonable load, tbe "neck size" n cannot be too small; especially, it follows from the Erlang formula (1) that n must exceed $(1 - \sqrt{1-p})\,N$ for line usage $p = m/N$. Similar lower bounds involving also the required loss can be derived. These lower bounds put a limit on how much one can concentrate (measuring global concentration by neck size), and they lead to $N \log N$ lower bounds for X even when there is not full access. For example, any network arranged in stages has

$$X \ge \frac{1}{2} e(1 - \sqrt{1-p})\, N \log N + 0(N)$$

if each customer's line carries p erlangs. Thus any sequence of networks that grow in the strong sense that p is bounded away from zero must grow like $N \log N$, if they are arranged in stages.

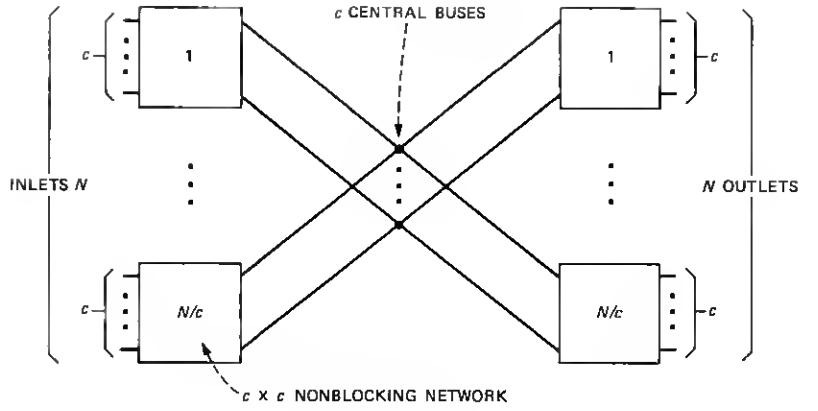

Fig. 2—Central bus network.

3.6 Asymptotic results

Three network structures lead, in the model to be used, to statistical equilibrium equations that have the "product form" solution, and so all their interesting parameters can be calculated exactly from a partition function, and tbeir behavior in the limit $N \to \infty$, $a = \lambda N^2 =$ constant, studied. They are (see Figs 2 througb 4.)

(*i*) The central bus concept—two large N-to-c nonblocking concentrators back to back, with c central buses

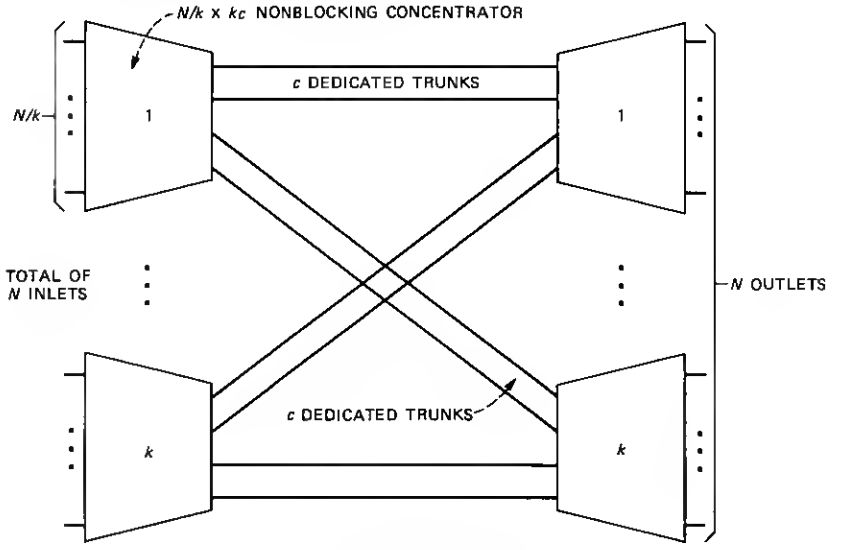

Fig. 3—Frame network.

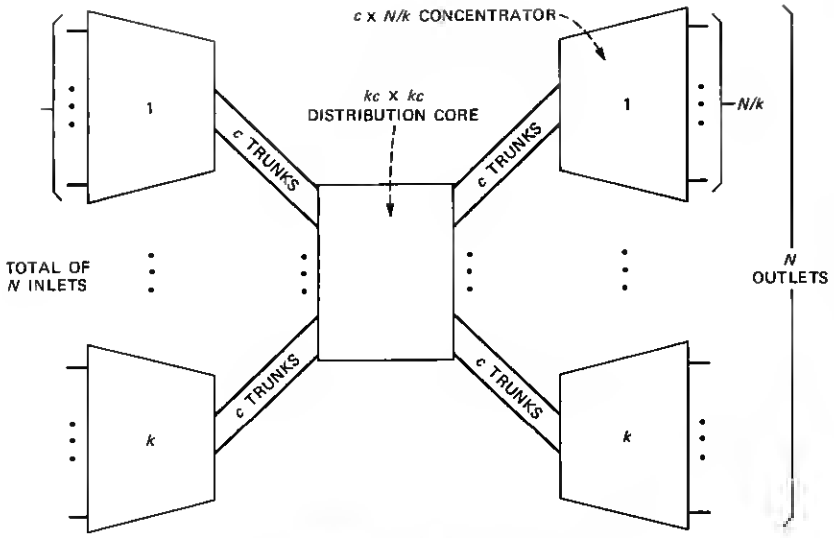

Fig. 4—Remote concentrator network.

(*ii*) The frame concept—large concentrators connected pairwise hy dedicated groups of c trunks

(*iii*) The remote concentrator concept—large concentrators, each connected to a central distrihution core hy its own group of c trunks.

In the "weak" limit as $N \to \infty$, $a = \lambda N^2 =$ constant, the first two structures are (not surprisingly) closely connected to Erlang's formula for loss: their prohabilities of loss are bounded above hy, and approach, the value $E(c, a)$. The third is more subtle and elusive, since loss can occur in either one of two relevant trunk groups, hut also more interesting. We can find $E(\cdot, \cdot)$-type bounds on the loss, hut the question is whether asymptotically

$$\text{loss} \leq 1 - (1 - b)^2 + o(1),$$

where b is the chance that all c trunks on one (any) concentrator are husy, in the limit, remains open. This quadratic upper bound is what the loss would he in the limit if the two relevant trunk groups were independent, each with "blocking" b. Very few of the many "hlocking polynomial" approximate formulas used in practice have heen vindicated hy so much as an inequality proved in a dynamical model.

It follows from our analyses that for each $\epsilon > 0$, and each of the three kinds of network structure considered, there exist arhitrarily large networks of that kind, with loss less than ϵ, and total offered traffic $a = \lambda N^2 =$ constant, whose crosspoint count X grows at most *linearly* with N. We view this as a limit result in the "weak" direction, since a constant amount of traffic $a = \lambda N^2$ is being divided up among

N lines, $N \to \infty$. The carried loads converge, and so the usage $p = m/N$, the erlangs carried per line, goes to zero. This linear growth is not inconsistent with the N log N orders mentioned earlier; the latter follow from combinatorial properties, or are the result of a different "limiting direction" in which p is bounded below away from zero. The above weak limit result is inappropriate for cases in which, as the network grows, each new customer is to supply a fixed number p_0 of added erlangs carried; this latter situation enjoins N log N growth. All the results about linear growth are given rigorous proofs, for the Markov process models adopted, by a passage from finite to infinite sources. In particular, no independence or other ad hoc assumptions are made to simplify the blocking estimate.

IV. COMPROMISES AND TRADE-OFFS

No matter what technology is used to build it, whether Strowger switches, crossbars, solid-state crosspoints, or time-division, the design of a telephone connecting network is inevitably a compromise between the competing criteria of cost and performance. Restricting attention to the traffic and operational aspects, it is nearly a truism that an overengineered network rich in switches will given unblemished service at an unacceptable cost, while one meagerly endowed with switches can only provide poor service at bargain cost. The trick the switching engineer must perform is to come up with designs that avoid these naive extremes.

The engineer's task is also affected by more subtle considerations, such as the following: the same pool of switching gear can be organized in an efficient way that is combinatorially optimal for connecting many pairs of customers in different patterns, i.e., realizing many assignments readily. Unfortunately, these efficient ways involve many stages and so are usually very complex and difficult to control, because putting up a call or removing one requires a lot of information and many decisions. Or the same gear can be hooked up in an inefficient, simple network of a few stages, which is easy to operate, but since it lacks combinatorial power it will have noticeably higher loss. Examples can be found in systems in which equipment is dedicated to handle certain geographically defined kinds of traffic (see Fig. 3).

Obviously, then, there are many trade-offs available to the designer between complexity, cost, control, and the various performance parameters, such as load and loss. Part of what traffic theory must provide, as a proper theoretical underpinning to network design, is an account of the options that are available (or unavailable) in the way of equipment, load, control and structural complexity, growth, and incurred loss. Such an account would describe the achievable regions in parameter space, some of the outer limits of possible designs, and the

achievable rates of growth of various indices of performance and complexity.

In the past such information for the large networks, which are of chief interest, has been sparse and difficult to obtain by means other than simulation and admittedly fuzzy theory. Thus, it is pertinent to have an account that retains its accuracy and force as the size of the network increases without bound. To give a specific example of this kind of information, we can put the following questions: For x, p, and ϵ specified positive numbers, are there arbitrarily large networks that have blocking probability at most ϵ, line usage at least p, and a number of crosspoints per terminal at most x? (To define blocking, let us assume for definiteness that call arrivals are random from finite sources by pairs, and holding times exponential—a "usual" model, that no one can quarrel with.) If there are such networks, how complex are they? That is, how fast does their complexity grow as measured by number of stages, amount of data needed to select a route, etc.?

V. THE VALUE OF BLOCKING

It is widely believed among telephone switching engineers that a positive probability of blocking is worth a great deal of switching and control equipment. Probably for this good reason alone, the famous nonblocking networks first invented by Charles Clos[4] have never been utilized on any but a small scale. Put another way, the canard says that to eliminate the last little bit of blocking will take an inordinate increase in equipment. But precisely what does the word "inordinate" mean here? Can we replace it by a mathematical function $E(b)$, which increases as the blocking b decreases, and whose interpretation is somehow that to achieve blocking b you need at least $E(b)$ in equipment? And how is $E(b)$ related to the structure of the network, to its size measured by number N of terminals on a side?

The remark that started the preceding paragraph is really only the tip of the iceberg: it is not exaggerating to claim that the mathematical basis of traffic engineering is the observation that very likely only a moderate number of telephone customers will want to talk to each other at the same time. The engineer provides enough lines, junctors, switches, trunks, etc., to take care of this overwhelmingly probable case, plus maybe a little extra for hedging his bets. The traffic theorist provides probabilistic models that make precise the meanings of "average," "likely," and "probable" in this setting. Of course the loads and losses the engineer seeks or achieves are subject to correction from customers' complaints, public commissions, and supervisors. But no matter what the numbers may be, the principle of not having to provide for the very unlikely events is universally accepted by all of those concerned. This principle amounts to public acceptance of pos-

itive blocking probability for public telephony; it stops designs from passing the inevitable knees in the cost curves, beyond which costs grow very fast for each increment in desired performance.

Naturally, then, we are interested in quantitative and mathematical expressions of the above principle. One can sensibly ask how much switching equipment can actually be saved by allowing a blocking probability of $\epsilon > 0$. Where in the system can you (and should you) save it? In particular, how is allowing a blocking of ϵ related to the rate of growth of the switching gear incurring that blocking as the number N of terminals gets large? It is known[5] that if $\epsilon = 0$ the needed gear in crosspoints is bounded above and below by

$$\text{const. } N \log N.$$

These results are purely combinatorial, and involve neither probability models nor traffic parameters. But is $N \log N$ still the right order of growth when $\epsilon > 0$, and the problem is posed in the context of our "usual" model, with a traffic parameter entering, such as the calling rate λ per idle inlet-outlet pair? The answer depends on just how λ varies relative to N as $N \to \infty$. If it is constant or decreases so slowly that the usage p is bounded away from zero, then $N \log N$ growth is necessary; if λN^2 remains constant, then for any $\epsilon > 0$, X need grow only linearly with N.

VI. THE KINDS AND ORIGINS OF LOSS

When we think about the probability of blocking in a network, it is useful to ask where and how it originates. At high levels of offered load, most of the loss incurred might be due primarily to frequent outright overloads of critical "bottlenecks," even without the occurrence of any combinatorial niceties such as mismatch. At lower levels of offered load, the fraction of loss owing to overloads is very small because the high occupancy states have very small probability, and most of the loss is due to mismatch in states of moderate occupancy and high total probability. It is not always possible to draw this distinction exactly in practice, nor is it necessary. It is important for theoretical purposes, though, because it suggests exactly analyzable large network structures (and models for them), in which the blocking is either all overload, or overload with certain simple kinds of mismatch. Such models can be used to study the trade-offs among the traffic and growth parameters; several are described in the latter half of this work.

VII. THE VIRTUE OF CONCENTRATION

Let it be agreed that for many terminals N and small calling rate λ per idle terminal-pair, the chance that many customers will want to

talk simultaneously is very small; if the principle of not providing for the very unlikely event is to be taken seriously, how should switching equipment be arranged? A standard and traditional method is to *concentrate* traffic.

The most efficient networks known today concentrate traffic from very many lightly loaded terminals into a heavily loaded central distribution core, in which link occupancy may run as high as 0.8 to 0.9, and then expand it back out in an inverse manner. The number of terminals entering the core is typically much smaller than the number of inlets, a feature that leads to a natural bottleneck or what we later call a "neck size." The blocking incurred can be thought of as being of two kinds, or arising from two sources: concentrator blocking, the inability of free inlets or outlets to get a line to the distribution core, and internal blocking in the core itself. Each of these sources of loss may in turn be due mostly to overload or to mismatch.

Thus, to reap the economic and operational values consequent on allowing blocking, the possible or maximum numbers of calls in progress at various places in the network are intentionally limited so as to save switching gear, the argument being as before that while having more calls in progress at these places is possible, it is so (or sufficiently) unlikely that there is no point in providing for it. We ask, how effectively can such limits curb the growth without impairing service?

Now the known nonblocking networks achieve their perfect operation by systematic *expansion*, the provision of more paths than can actually be used at one time; this is the antithesis of the concentrator-cum-distribution core idea usually used in practice. These nonblocking networks exhibit $N \log N$ growth, as they must. Since concentration is the opposite of expansion, it should lead to a saving in crosspoints. How big a saving is it? Especially, when can concentration reduce the order of growth to something slower than $N \log N$?

VIII. ASYMPTOTICS FOR LARGE NETWORKS

In seeking to answer some of the preceding questions, we shall examine the behavior of network parameters as the number N of terminals on a side becomes arbitrarily large. To fix ideas we begin with some thought-experiments that will lead to more specific questions.

Accepting for a moment the conventional wisdom that all efficient large networks use concentrating switches, we imagine a sequence of such networks with more and more terminals, and ask: What can happen to the efficiency of these networks as they grow? Is it possible to keep the loss below a specified amount without having the crosspoint count or the number of stages grow very fast? Are there large networks which, though they may not be in the running as red-hot field designs

for a particular technology, nevertheless are of interest because their load, loss, and cost can be easily and accurately calculated or estimated? If there are such networks, what combinatorial features account for the ease of calculation? Where are "most" of the crosspoints, in the concentration stages, or in the distribution core?

IX. LIMIT DIRECTIONS

Needless to say, some care must be taken in carrying out the analyses needed to answer these questions. For most purposes, it is enough to make more exact the way in which the traffic and performance parameters are to vary as N grows; usually some group or function of them is constrained to stay in a given set. Such constraints define "directions" in which limits or other asymptotics are being sought, and they provide useful ways of looking at the performance of very large networks. Two such directions, leading to very different growth rates, will be of interest here.

9.1 The "weak" limit

For example, we can let the offered load per idle pair λ get small and N get big, so that $a = \lambda N^2$ is a fixed constant. This amounts to letting the process of attempted calls become Poisson with rate a; the corresponding limit process is sometimes called a "passage from finite to infinite sources," and in traffic theory is often associated with familiar notions such as the Poisson approximation to the binomial distribution. We shall show by examples that some interesting limits of this kind exist and can be evaluated, leading to functions and concepts well-known in traffic theory, such as the Erlang loss $E(c, a)$. Such calculations lead to information about the growth rate of cost and complexity for large networks that have specified load and loss.

9.2 The strong constraint

A very different condition, to be used later, is that the usage $p = m/N$ of our sequence of growing networks be bounded away from zero. It says roughly that each new customer, as N grows, adds a fixed amount of carried load to the network, at least. This condition, incompatible with the "weak" limit, leads to $N \log N$ growth in crosspoints, and is not, as far as we know, associated with any limits in distribution, the way the weak limit is. That is why we call it a condition and not a limit. It is physically natural for networks to grow in this manner, and so this is an important condition to consider.

X. NETWORK STATES

We shall use a model for the structural and combinatorial aspects of

a connecting network. This model arises by considering the network structure to be given by a graph G whose vertices are the terminals of the network, and whose edges represent crosspoints between terminals by pairs, with some of the terminals designated as inlets or outlets. Calls in the network are described by paths on G from an inlet to an outlet. Thus, a connecting network ν is a quadruple $\nu = (G, I, \Omega, S)$, where G is a graph depicting networking structure, I is the set of vertices of G which are inlets, Ω is the set of outlets, and S is the set of permitted or physically meaningful states. It is possible that $I = \Omega$ (one-sided network), that $I \cap \Omega = \phi$ (two-sided network), or that some intermediate condition obtain, depending on the "community of interest" aspects of the network v. Variables w, x, y, and z at the end of the alphabet denote states, while u and v denote a typical inlet and a typical outlet, respectively.

A possible state x can be thought of as a set of disjoint chains on G, each joining I to Ω. Not every such set of chains need represent a state in S: wastefully circuitous chains may be excluded from S. The set S is partially ordered by inclusion $\leq$, where $x \leq y$ means that state x can be obtained from state y by removing zero or more calls. It is reasonable that if y is a state and x results from y by removal of some chains, then x should be a state too, i.e., S should be closed under "hangups." It can be seen from this requirement that the set S of permitted states has the structure of a semilattice, that is, a partially ordered system whose order relation is definable in terms of a binary operation $\cap$ that is idempotent, commutative, and associative, by the formula $x \leq y$ iff $x = x \cap y$. Here for $x \cap y$ we can simply use literal set intersections: $x \cap y$ is exactly the state consisting of those calls and their respective routes that are common to x and y.

An *assignment* is a specification of what inlets are to be connected to what outlets. The set A of assignments can be represented as the set of all fixed-point-free correspondences from subsets of I to Ω. The assignments form a semilattice in the same way that the states do, and A is related to S as follows: call two states x, y in S equivalent as to assignment, written $x \sim y$, iff all and only those inlets $u \in I$ are connected in x to outlets $v \in \Omega$, which are connected to the same v in y, though possibly by different routes. The realizable assignments can then be identified with the equivalence classes of states under $\sim$, and there is a natural map $\gamma: S \to A$, the projection that carries each state x into the assignment $\gamma(x)$ it realizes, i.e., the equivalence class it belongs to under $\sim$.

With x and y states such that $x \geq y$, it is convenient to use $x - y$ to mean the state resulting from x by removing from x all the calls in y. Similarly, with a and b assignments such that $a \geq b$, we use $a - b$ to mean the assignment resulting from a by dropping all the connections

intended in b. Note that here $x - y$, $a - b$ bave their usual set-theoretic meaning.

It can now be seen that the map γ is a semilattice homomorphism of S into A, with the properties:

$$x \geq y \Rightarrow \gamma(x) \geq \gamma(y)$$

$$x \geq y \Rightarrow \gamma(x - y) = \gamma(x) - \gamma(y)$$

$$\gamma(x \cap y) \leq \gamma(x) \cap \gamma(y)$$

$$\gamma(x) = \phi \Rightarrow x = 0 = \text{zero state, with no calls up.}$$

Not every assignment need be realizable by some state of S. Indeed, it is common for practical networks to realize only a vanishing fraction of the possible assignments, and tbe networks that do realize every assignment, the so-called *rearrangeable* networks, have been the objects of substantial theoretical study. Thus, the image set $\gamma(S)$ of realizable assignments is typically much smaller than the set A in which it is embedded. A *unit* assignment is, naturally, one that assigns exactly one outlet to some inlet, and it corresponds to having just one call in progress. It is convenient to identify calls c and unit assignments, and to write $\gamma(x) \cup c$ for the larger assignment consisting of $\gamma(x)$ and the call c together, with the understanding, of course, that c is "new in x" in the sense that neither of its terminals is busy in x.

We denote by A_x the set of states that are immediately above x in the partial ordering $\leq$ of S, and by B_x the set of those that are immediately below. Thus,

$$A_x = \{\text{states reachable from } x \text{ by adding a call}\}$$
$$B_x = \{\text{states reachable from } x \text{ by bangup}\}.$$

For c new in x, let $A_{cx} = A_x \cap \gamma^{-1}[\gamma(x) \cup c]$; A_{cx} is the subset of states of A_x that could result from x by putting up the call c, because $\gamma^{-1}\gamma(y)$ is precisely the equivalence class of y under $\sim$. If A_{cx} is empty then we say c is blocked in x: there is no $y \in A_x$ that realizes the larger assignment $\gamma(x) \cup c$. It can be seen that with F_x the set of new calls of x that are not blocked, the family $\{A_{cx}, c \in F_x\}$ forms the partition of A_x induced by equivalence $\sim$.

XI. ROUTING OF CALLS

We shall use a routing matrix $R = (r_{xy})$ as a convenient formal description of how routes are chosen for calls. The class of routing matrices, R, can be described thus: for each $x \in S$ let Π_x be the partition of A_x induced by the relation $\sim$ of "having the same calls up", or satisfying the same assignment of inlets to outlets; it can be seen that Π_x consists of exactly the sets A_{cx} for c free and not blocked

in x; for $Y \in \Pi_x$, r_{xy} for $y \in Y$ is to be a probability distribution over Y, tbat is $r_{xx} \geq 0$ and $\sum_{y \in Y} r_{xy} = 1$; r_{xy} is to be 0 in all other cases.

The interpretation of the routing matrix as a method of choice is to be this: any $Y \in \Pi_x$ represents all the ways in which a particular call c (free and not blocked in x) co*uld* be completed when the network is in state x; for $y \in Y$, r_{xy} is the cbance (or fraction of times) that if call c arises in state x it will be completed by being routed in the network so as to take the system to state y. The distribution $\{r_{xy}, y \in Y\}$ indicates how the calling rate owing to c is to be spread over the possible ways of putting up this call. Evidently, such a description of routing could be made time-dependent, and extended to cover refusal of unblocked calls as an option; we do not consider these possibilities here. The problem of choosing an optimal routing matrix R bas been worked on at some lengtb.

XII. STOCHASTIC MODEL

We now recall[2] a stocbastic model for the traffic offered to a network. A Markov stochastic process x_t taking values on S can be based on these simple probabilistic and operational assumptions:

(*i*) Holding times of calls are mutually independent variates, eacb with tbe negative exponential distribution of unit mean.

(*ii*) If u is an inlet idle in state $x \in S$, and $v \neq u$ is any outlet, there is a conditional probability $\lambda h + o(h)$, $\lambda > 0$, as $h \to 0$, that u attempts a call to v in $(t, t + h)$ if $x_t = x$.

(*iii*) A routing matrix $R = (r_{xy})$ is used to choose routes, as follows: If c $= \{(u, v)\}$ is a call free and not blocked in x, then the fraction of times tbat the system passes from x to $y \in A_{cx}$ if c arises when $x_t = x$ is just r_{xy}.

(*iv*) Blocked calls and calls to busy terminals are declined, with no change of state.

It is convenient to collect these assumptions into a transition rate matrix $Q = (q_{xy})$, the generator of x_t; this matrix is given by

$$q_{xy} = \begin{cases} 1 \quad \text{if } y \in B_x \\ \lambda r_{xy} \text{ if } y \in A_x \\ -|x| - \lambda s(x) \text{ if } y = x, \text{ with } s(x) = |F_x| \\ 0 \text{ otherwise,} \end{cases}$$

and the associated statistical equilibrium (or state) equations take the simple form

$$[|x| + \lambda s(x)]p_x = \sum_{y \in A_x} p_y + \lambda \sum_{y \in B_x} p_y r_{yx} \; x \in S,$$

where $\{p_x, x \in S\}$ is the asymptotic distribution of x_t. Here $|x|$ denotes the number of calls in progress in x, and $s(x)$ is the number of

unblocked idle inlet-outlet pairs in x, the possible "successes" in x; note that $s(x) = |F_x|$.

XIII. PARAMETERS OF INTEREST FOR DESIGN AND ENGINEERING

We shall frequently use about a dozen basic parameters, characteristic of the operating network, and important for these reasons: They describe load, cost, and performance, or they can be measured readily, or they arise naturally in the associated traffic theory and are convenient for calculations and asymptotic analyses. For two-sided networks ν, these parameters are the following:

λ = calling rate per idle inlet-outlet pair
N = number of terminals (inlets, or outlets) on each side
bl = probability of blocking
m = carried load = expected number of calls in progress
p = usage = m/N = erlangs carried per terminal
σ = standard deviation of number of calls in progress
X = total number of crosspoints
s = number of stages (if ν is arranged in stages)
$a = \lambda N^2$ = total offered load when everyone is idle
$w = \max_{x \in S} |x|$ = maximum possible number of calls in progress
n = "neck size," defined for ν arranged in stages separating junctor groups $T_1, T_2, \cdots T_{s+1}$, as the cardinality of the smallest T_i.

Remark: The parameter $\alpha = \lambda N^2$ is a convenient abbreviation for total offered load, especially for certain weak "large network" asymptotics for which λN^2 is held constant as $\lambda \to 0$ and $N \to \infty$.

Remark: The ratios w/N and n/N are rough *global* measures of concentration, global because there could be, for example, remote local concentrators with a concentration ratio different from each of these. Clearly, $w \leq n$, when both w and n are defined.

Notation: We write $X(\nu)$, $p(\nu)$, etc, whenever it is necessary to express the dependence of a parameter on the network ν.

XIV. THE PARAMETER SURFACE

In the early[1] applications of traffic theory to trunking problems, a central role was played by Erlang's loss formula, which depended on two parameters, the load a, and the number c of trunks. For connecting network studies, though, to take into account at least the size of the network and the "finite source" effect, if not other network features, a modification of Erlang's formula is more suitable. (The finite source effect is a recognition that busy terminals generate no traffic.) Such a formula has been derived in earlier work.[2] We shall exhibit many

useful results that follow from it, or from it together with reasonable but special hypotheses, such as the property of a network that it is made of square switches, or is arranged in stages, or provides full access.

For a two-sided network witb N terminals on a side, the load, loss, load deviation, and rate parameter, λ, are related by the following formula:

Generalization of Erlang's formula to networks:

$$1 - bl = \frac{1}{\lambda} \frac{m}{(N-m)^2 + \sigma^2} \tag{1}$$

$$\alpha(1 - bl) = \frac{m}{(1-p)^2 + (\sigma/N)^2}.$$

Proof: $1 - bl$ is the fraction of attempted calls that are not blocked. By the law of large numbers, this fraction is the rate of successful attempts divided by the total rate of attempts, both in equilibrium. For a state x let $s(x)$ be the number of inlet-outlet pairs (u, v) that are not blocked in x; "$s(x)$" stands for "successes in x." The success rate is then $\lambda \sum_{x \in S} p_x s(x)$ and the total attempt rate $\lambda \sum_{x \in S} p_x (N - |x|)^2$. However, in equilibrium, the rate in equals the rate out, so

$$\lambda \sum_{x \in S} p_x s(x) = \sum_{x \in S} p_x |x| = m.$$

Evidently the total attempt rate can be written as $\lambda[(N-m)^2 + \sigma^2]$, and the general Erlang formula is proved.

Remarks: The gist of the Erlang formula is that six of the parameters of interest cannot assume arbitrary values but must lie on a surface described by a simple rational function. It is apparent that similar but more complex formulas can be proved for one-sided networks, or two-sided ones with different numbers of terminals on each side; we shall not consider these, because the basic ideas are the same. Note that blocking, a complicated quantity in the model, is determined solely by the first two moments of the number of calls in progress. Also, if N and λ vary so that $(1-p)^2 + (\sigma/N)^2$ approaches 1, the formula approaches the exact form $m = a(1 - bl)$ it has in the "true" Erlang case.[1]

XV. SOME TECHNICAL RESULTS

The generalized Erlang formula has many useful consequences. Some of them are summarized in the next few results, all of which are additional relationships among the engineering parameters.

Lemma: $\dfrac{\sigma^2}{N^2} < p - p^2.$ (2)

Proof: Clearly, $E|x_t|^2 = \sum_{x \in S} p_x |x|^2 \geq Nm$. Hence,

$$\sigma^2 = E|x_t|^2 - (E|x_t|)^2 \leq Nm - m^2.$$

Lemma: $\dfrac{m}{a(1 - bl)} \leq 1 - p.$ (3)

Proof: Lemma (2) implies that

$$p^2 - 2p + 1 + \frac{\sigma^2}{N^2} < 1 - p.$$

By (1) the left-hand side is $m/a\ (1 - bl)$.
Representation of N:

$$N = \frac{m + \sqrt{m^2 - \left[1 - \dfrac{m}{a(1 - bl)}\right](m^2 + \sigma^2)}}{1 - \dfrac{m}{a(1 - bl)}}$$

$$= \frac{m(1 + \sqrt{1 - p})}{1 - \dfrac{m}{a(1 - bl)}}, \tag{4}$$

where

$$p = \left[1 - \frac{m}{a(1 - bl)}\right]\left[1 + \frac{\sigma^2}{m^2}\right]$$

and

$$0 < p < 1.$$

Proof: N is one of the quadratic roots

$$\frac{m(1 \pm \sqrt{1 - p})}{1 - \dfrac{m}{a(1 - bl)}}. \tag{5}$$

To show that the plus branch is the right one we note that the quadratic (in y) function

$$y^2\left[1 - \frac{m}{a(1 - bl)}\right] - 2my + m^2 + \sigma^2 \tag{6}$$

equals $m^2 + \sigma^2$ at $y = 0$, and has a negative minimum at

$$y^* = \frac{m}{1 - \dfrac{m}{a(1-bl)}}.$$

This value is a minimum because the second derivative is

$$2\left[1 - \frac{m}{a(1-bl)}\right] > 2p$$

by Lemma (3). This value is negative because it is

$$m^2 + \sigma^2 - \frac{m^2}{1 - \dfrac{m}{a(1-bl)}} \tag{7}$$

and tbe Erlang formula (1) implies that

$$\sigma^2 + m^2 = 2mN - N^2\left[1 - \frac{m}{a(1-bl)}\right]. \tag{8}$$

Substituting this into (7) we see that the value is

$$-N^2\left[1 - \frac{m}{a(1-bl)}\right] + 2mN - \frac{m^2}{1 - \dfrac{m}{a(1-bl)}},$$

clearly negative. Thus y^* separates the two real roots of (6). However, Lemma (3) implies that $N > y^*$, so N must be given by tbe plus sign in (5). It is obvious that p is positive; to show $p < 1$ we divide (8) by m^2 and multiply by $1 - \dfrac{m}{a(1-bl)}$ to get

$$p = 1 - \left\{1 - \frac{N}{m}\left[1 - \frac{m}{a(1-bl)}\right]\right\}^2 < 1,$$

which incidentally strengthens Lemma (3) to

$$p < 1 - \frac{m}{a(1-bl)} < 2p.$$

Lemma: If ν is such that, at most, w calls can be in progress, then for $w < N$

$$(i)\ \ bl \geq 1 - \frac{m}{\lambda(N-w)^2}$$

$$(ii)\ \ w \geq N(1 - \sqrt{1-p})$$

$$(iii)\ \ p \leq \frac{2w}{N} - \left\{\frac{w}{N}\right\}^2. \tag{9}$$

Proof: In this situation the average calling rate is greater than or equal

to $(N - w)^2\lambda$, so (*i*) follows from the generalized Erlang formula (1). Thus also

$$\frac{m}{\lambda N^2[(1-p)^2 + \sigma^2 N^{-2}]} \leq \frac{m}{\lambda(N-w)^2}(N-w)^2$$

$$\leq N^2[(1-p)^2 + \sigma^2 N^{-2}] = N^2 \frac{m}{a(1-bl)}.$$

But by Lemma (3),

$$\frac{m}{a(1-bl)} \leq 1 - p$$

and thus (*ii*) and (*iii*) via

$$N - w \leq N\sqrt{1-p}.$$

The same argument proves

$$w \geq N[1 - \sqrt{(1-p)^2 + \sigma^2 N^{-2}}].$$

XVI. NONCONCENTRATING NETWORKS GROW LIKE *N* LOG *N* EVEN IN THE WEAK LIMIT

One way to display the value of concentrating traffic is to show how bad things are when you do not do it. We shall look at a large class of practical networks that have no concentration, and show that to all intents and purposes, these networks grow like $N \log N$ for N terminals. In particular, without concentration, these networks have no way of trading off nonzero or even substantial blocking (up to some $\epsilon > 0$) for slow growth, i.e., slower than $N \log N$, or constant $\times N \log N$ with a constant that depends on usage p. Especially, these networks have $N \log N$ growth even in the weak limit $a = \lambda N^2 =$ constant; we show later that, in this special case, linear growth is achievable if concentration is used, even with blocking kept less than ϵ.

Now the prototypical network without concentration is one that is made of *square* switches arranged in stages, and these constitute the class we consider. Since engineers are primarily interested in networks with a high degree of symmetry, which "look the same" from any terminal within a junctor group, or from any inlet, or outlet, we shall restrict attention to networks ν with these properties:

(*i*) ν has $s \geq 1$ stages

(*ii*) The switches in each stage are square, and identical.

Note that the number of stages is not fixed, and that different stages may have different numbers of switches; also, the interconnection patterns between the stages (in old terminology, the *cross-connect fields*) can represent arbitrary N-permutations.

Theorem: Let $n_1, n_2, \cdots, n_s$ be s divisors of N, possibly with repetitions, and consider a network ν with N/n_i $n_i \times n_i$ switches in its ith stage, and blocking $bl < \epsilon$. Then its crosspoint count $X(\nu)$ satisfies

$$X(\nu) \geqq N \sum_{i=1}^{s} n_i \geq N \log N + N(s + \log(1 - \epsilon)) - a. \tag{10}$$

Remark: The number of outlets reacbable from an inlet is at most $\Pi_{i=1}^{s} n_i$. If every inlet can reach every outlet, then this product exceeds $N - 1$, and in this "usual" case

$$\sum_{i=1}^{s} n_i \geq s + \log \prod_{i=1}^{s} n_i \geq s + \log N \tag{11}$$

and the conclusion of the theorem follows. In the opposite "unusual" case, $\Pi_{i=1}^{s} n_i < N$, some calls are permanently blocked, and the network does not provide full access. Thus, one point of the theorem is that even in this combinatorially poor situation, blocking cannot save more than a linearly growing number of crosspoints when $a = \lambda N^2$ is constant or grows at most linearly, so that $X(\nu) = O(N \log N)$.

Proof: Only the "unusual" case $N > \Pi_{i=1}^{s} n_i$ need be considered. Let $s(x)$ be the number of possible "successes" in x, i.e., of inlet-outlet pairs idle and not blocked in state x, so that

$$(N - |x|)^2 - s(x)$$

is the number of idle inlet-outlet pairs that cannot be connected in x. In the unusual case an idle inlet cannot reach at least $N - \Pi_{i=1}^{s} n_i$ outlets. In state x, at most $|x|$ of these unreachable outlets are busy; thus there are at least

$$N - \prod_{i=1}^{s} n_i - |x|$$

idle outlets with no path to our test inlet. This being true for each idle inlet, and there being $N - |x|$ idle inlets, the number β_x of blocked idle inlet-outlet pairs in state x satisfies

$$\beta_x \geq (N - |x|) \left(N - \prod_{i=1}^{s} n_i - |x| \right),$$

or since $(N - |x|)^2 = s(x) + \beta_x$,

$$s(x) \leq (N - |x|) \prod_{i=1}^{s} n_i.$$

Averaging this inequality with respect to the equilibrium state probabilities $\{p_x, x \in S\}$, and noting that $m = \sum_{x \in S} |x| p_x = \lambda \sum_{x \in S} s(x) p_x$, we find

$$\frac{m}{\lambda} \leq (N-m) \prod_{i=1}^{s} n_i$$

$$\prod_{i=1}^{s} n_i \geq \frac{Nm}{a(1-p)}.$$

Therefore,

$$\begin{aligned} X(\nu) &= N \sum_{i=1}^{s} n_i \\ &\geq N\left(s + \log \prod_{i=1}^{s} n_i\right) \\ &\geq Ns + N \log N + N \log \frac{m}{a(1-p)}. \end{aligned}$$

To find a suitable lower bound on the last term we use the basic generalization (1) of Erlang's formula:

$$\begin{aligned} 1 - bl &= \frac{m}{\lambda(N-m)^2 + \lambda\sigma^2} = \frac{m}{a[(1-p)^2 + \sigma^2 N^{-2}]} \\ &\leq \frac{m}{a(1-p)^2}. \end{aligned}$$

Since the blocking bl is less than ϵ, one finds

$$\log(1-\epsilon) \leq \log(1-bl) \leq \log \frac{m}{a(1-p)} - \log(1-p)$$

$$X(\nu) \geq N \log N + sN + N \log(1-p) + N \log(1-\epsilon).$$

Next we argue that, as in Lemma (2),

$$\sigma^2 N^{-2} \leq p - p^2$$

$$(1-p)^2 + \sigma^2 N^{-2} \leq 1 - p,$$

and so by the Erlang formula (1) again,

$$p = \frac{1-bl}{N} a[(1-p)^2 + \sigma^2 N^{-2}] \leq \frac{1-p}{N} a \leq \frac{a}{N+a}$$

whence

$$\log(1-p) > \log N - \log(N+a) \geq -\frac{a}{N}$$

and the proof is complete.

XVII. GROWTH OF NETWORKS ARRANGED IN STAGES

Almost all the connecting networks used in practice are *made of*

stages, or *arranged in stages*. This property means that the terminals of the network are partitioned into disjoint sets $T_1, T_2, \cdots, T_{s+1}$, whicb are simply ordered as indicated; the first set T_1 consists of tbe inlets, the next $s - 2$ sets are internal junctors, and the last set T_{s+1} consists of the outlets; crosspoints are placed only from terminals in a given set T_i to terminals in the next set T_{i+1} in tbe ordering (see Fig. 1). Thus every call in progress is a path from an inlet to an outlet that passes through each set T_i once in the order specified. The crosspoint pattern between successive sets is called a stage of switching, and is representable as a bipartite graph. The sets T_i need not all have the same numbers of terminals; if they do not, there is expansion or concentration. The size of tbe smallest T_i is called tbe "neck size," and is of course an upper bound on the number of calls in progress:

$$|x| \leq n = \min_{1 \leq i \leq s+1} |T_i|.$$

We shall restrict attention to *symmetric* networks, in which the network looks the same to every terminal in a given T_i, $i = 1, \cdots, s + 1$. A network provides *full access* if every inlet-outlet pair can be connected by a patb through the network, with no doubling back allowed. Full access is a natural convenient condition that greatly simplifies arguments, but it is not necessary for proving $N \log N$ growth.

Theorem: Let ν be a symmetric network with N inlets (& outlets), with neck size n, providing full access through s stages. Then the crosspoint count $X(\nu)$ satisfies

$$X(\nu) \geq en \log N \quad e = 2.71828 \cdots . \tag{12}$$

Proof: Let n_i, $i = 1, \cdots, s$, be tbe number of crosspoints in stage i connected to a junctor between stages i and $i + 1$. By symmetry this number is the same for all such junctors or terminals. If stage i is represented by a bipartite graph, n_i is the degree of each "input" vertex. Thus, if the neck size is n, tben stage i has at least nn_i crosspoints, and so

$$X(\nu) \geq n \sum_{i=1}^{s} n_i.$$

Since ν provides full access it must be true that

$$N \leq \prod_{i=1}^{s} n_i;$$

for an inlet can reach no more than $\Pi_{i=1}^{s} n_i$ outlets; so if N exceeded this product there would be some it could not reach. Hence, by the inequality linking arithmetic and geometric means,

$$X(v) \geq n \sum_{i=1}^{s} n_i$$

$$\geq ns \left(\prod_{i=1}^{s} n_i \right)^{\frac{1}{s}}$$

$$\geq nsN^{\frac{1}{s}}.$$

But $sN^{\frac{1}{s}}$, viewed as a function of a real variable s, has a unique minimum at $s = \log N$, so that

$$X(v) \geq n \log N N^{\frac{1}{\log N}} = en \log N, \quad e = 2.71828 \cdots.$$

Remark: If the neck size is N, as it is for networks made of square switches, then for symmetric v providing full access

$$X(v) \geq e\, N \log N.$$

Lemma: If v has neck size $n < N$, then

$$n \geq N\left\{1 - \left[\frac{m}{a(1-bl)}\right]^{1/2}\right\} \geq N\{1 - \sqrt{1-p}\}. \tag{13}$$

This result links the neck size n to the performance parameters m and bl and to the traffic parameter a. The second inequality is remarkable in involving only the line usage p; the higher p is to be, the closer the neck size must be to N.

Proof: If v has neck size n, then at most n calls can be in progress at a time, so that $N - |x| \geq N - n$, and by the Erlang formula (1)

$$1 - bl = \frac{m}{\lambda \Sigma p_x (N - |x|)^2} \leq \frac{m}{\lambda (N-n)^2}$$

$$(N-n)^2 \leq N^2 \frac{m}{a(1-bl)}.$$

The second inequality follows from Lemma (3); it leads at once to this basic result:

Theorem: If v is a symmetric network arranged in stages, providing full access and with each inlet carrying p erlangs, then

$$X(v) \geq e(1 - \sqrt{1-p})N \log N. \tag{14}$$

Proof: (12) and (13).

Discussion: This inequality says that any symmetric network providing full access has const. $N \log N$ crosspoints, where the constant depends only on the line usage, no matter what blocking is incurred. Using the

first inequality in (13) gives a larger constant, now dependent on loads (offered and carried) and blocking.

Theorem: Let ν be a network arranged in stages, not necessarily providing full access. Then

$$X(\nu) \geq e(1 - \sqrt{1-p})[\tfrac{1}{2}N \log N - \tfrac{1}{2}N \log \lambda + N \log(1 - bl) + N \log p]. \quad (15)$$

Proof: As in the proof of Theorem (12) we find

$$X(\nu) \geq ns\left(\prod_{i=1}^{s} n_i\right)^{\frac{1}{s}}. \quad (16)$$

Next, the averaging argument of Theorem (14) gives

$$\prod_{i=1}^{s} n_i \geq \frac{Nm}{a(1-p)} = \frac{p}{\lambda(1-p)}.$$

For $b > 0$, $sb^{\frac{1}{s}}$ assumes a unique minimum at $s = \log b$, and $(b)^{\frac{1}{\log b}} = e = 2.71828, \cdots$, so by (16),

$$X(\nu) \geq ne \log \frac{p}{\lambda(1-p)}$$

$$\geq e(1 - \sqrt{1-p})N[\log p - \log \lambda - \log(1-p)]. \quad (17)$$

The generalized Erlang formula (1) can be put in the forms

$$\lambda N(1 - bl)[(1-p)^2 + \sigma^2 N^{-2}] = p \quad (18)$$

$$\lambda N(1 - bl)(1-p)^2 + (1-p) - 1 + \lambda N(1 - bl)\sigma^2 N^{-2} = 0. \quad (19)$$

The second form is a quadratic equation for $1 - p$ whose solution, picking the plus branch, is

$$1 - p = \frac{\sqrt{1 + 4\lambda N(1-bl)[1 - \lambda(1-bl)\sigma^2 N^{-1}]} - 1}{2\lambda N(1-bl)}.$$

By (18) above, the quantity (factor) $1 - \lambda(1 - bl)\sigma^2 N^{-1}$ under the square root is equal to

$$1 - p\frac{\sigma^2}{(N-m)^2 + \sigma^2}$$

and lies strictly between 0 and 1. Therefore,

$$1 - p < \frac{\sqrt{1 + 4\lambda N(1-bl)} - 1}{2\lambda N(1-bl)}.$$

Let $y = 2\lambda N(1 - bl)$ for short, so that

$$1 - p < \frac{\sqrt{1+2y} - 1}{y}.$$

Hence,

$$\log(1-p) < \log \frac{\sqrt{1+2y}-1}{y} = \log \frac{2}{\sqrt{1+2y}+1}$$

$$-\log(1-p) > \log(\sqrt{1+2y}+1) - \log 2$$

$$> \frac{1}{2}\log(1+2y) - \log 2$$

$$> \frac{1}{2}(\log 4 + \log \lambda + \log N + \log(1-bl)) - \log 2.$$

Returning now to formula (17) we find

$$X(\nu) \geq e(1-\sqrt{1-p})N[\log p - \tfrac{1}{2}\log \lambda + \tfrac{1}{2}\log N + \tfrac{1}{2}\log(1-bl)].$$

Remark: Theorem (15) shows that full access is not necessary for $N \log N$ growth; it just makes the constant bigger and the argument simpler.

Theorem: Let ν_N be a sequence of networks on N terminals arranged in stages, and such that

(*i*) $p(\nu_N) \geq p_0 > 0$
(*ii*) $bl(\nu_N) \leq \epsilon$ (20)
(*iii*) $\lambda(\nu_N)$ is bounded.

Then as $N \to \infty$

$$X(\nu_N) \geq \frac{e}{2}(1-\sqrt{1-p_0})N \log N + O(N).$$

Proof: Immediate from (15).

Remarks: Theorems (15) and (20) of course apply also to the networks made of stages of square switches considered in Theorems (10) and (11). However, it should be noted that the possible traffic asymptotics in the two theorems are different, although they might overlap. In (11) $a = \lambda N^2$ grows at most linearly, while in (20) it grows at least linearly; in (11) $a = \lambda N^2$ could be identically a constant (the weak limit case), so that λ and p both go to zero, and $X(\nu)$ grows like $N \log N$ instead of linearly as it might [see Theorem (24)]; in (20), on the other hand, p is bounded away from zero, hence λN is also, so $a = \lambda N^2$ increases at least linearly, and $X(\nu)$ grows like constant $\times N \log N$, with constant depending on the lower bound for p. The point is that the absence of concentration exemplified in the square switches compels $N \log N$ growth even in the weak limit ($a = \lambda N^2$ constant, p vanishing), while

in general if concentration is allowed it takes the "strong condition" $p \geq p_0 > 0$ to force $N \log N$ growth.

XVIII. ANALYZABLE LARGE NETWORKS

We turn now to the study of three simple patterns or structures for networks with concentration. Interest in them arises from the fact tbat their loads, losses, and complexity can be calculated or rigorously bounded for arbitrarily large values of N. Most of them embody features, such as frames and concentrators, whicb are familiar in telephone network design, and some provide tenuous links to previous approximate blocking formulas based on independence assumptions. These formulas suggest inequalities stating that certain natural "blocking polynomials" are in fact upper bounds on the probability of blocking; their proof or disproof has eluded us so far, but they are worth mentioning nevertheless.

XIX. CENTRAL BUSES CONCEPT

A useful extreme case is a network that, like the trunk group, has no blocking states until a certain number c of calls in progress is reached, at which point *all* calls are blocked. One way to build such a network is to concentrate both the N inlets and the N outlets down to c terminals in a nonblocking way, and then to put a c-by-c nonblocking network in between, as shown in Fig. 5. A better way results when we note that the central network is superfluous; all you need are c *central buses* with "expanding" networks on each side such that any idle terminal can reach an idle bus. An arrangement of this kind is shown in Fig. 2, in which each bus has an appearance on every (inlet or outlet) subnetwork; when these are nonblocking, a theoretically useful solvable case results. We call such networks c*entral bus networks.*

Remark: Clearly, the central bus idea for networks springs right out of the idea that in a large network with lightly loaded lines, only a

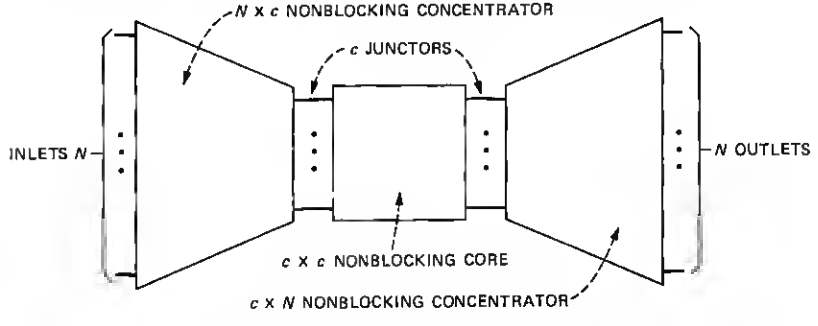

Fig. 5—Network nonblocking up to c calls.

moderate number of customers will be talking. The thought is this: if the low and moderate occupancy states have all the probability, let's use all our switching gear to make *them* as nonblocking as possible, and just ban the unlikely high occupancy states altogether. Accordingly, the designer guesses or calculates what that moderate occupancy might be, on the average, and provides some larger number c of central buses, with nonblocking access for everyone.

Remark: The central bus network is also a good candidate for the best disposition of a fixed number X of crosspoints at very low traffic λ. For it is known[2] that

$$\text{blocking} = \text{const.}\ \lambda^m + o(\lambda^m) \quad \text{as} \quad \lambda \downarrow 0,$$

where

$$m = \min_{x \in S} \{|x|: \text{some call is blocked in } x\}.$$

Thus, of all networks made out of X crosspoints, the ones for which m is largest will have asymptotically least blocking at low traffic, as $\lambda \to 0$.

For traffic purposes, the number of calls in progress is an adequate notion of state for central bus networks. Under our assumptions, the equilibrium probabilities p_n of n calls up satisfy

$$p_n = p_0 \frac{\lambda^n}{n!} (N - n + 1)^2 (N - n + 2)^2 \cdots N^2$$

and so the blocking is

$$bl = \frac{\dfrac{\lambda^c}{c!} (N - c)^2 (N - c + 1)^2 \cdots N^2}{N^2 + \displaystyle\sum_{j=1}^{c} \frac{\lambda^j}{j!} (N - j)^2 (N - j + 1)^2 \cdots N^2}.$$

We introduce the parameter $a = \lambda N^2$ by writing this as

$$bl = \frac{\dfrac{a^c}{c!} \left(1 - \dfrac{c}{N}\right)^2 \left(1 - \dfrac{c-1}{N}\right)^2 \cdots \left(1 - \dfrac{1}{N}\right)^2}{1 + \displaystyle\sum_{j=1}^{c} \frac{a^j}{j!} \left(1 - \frac{j}{N}\right)^2 \left(1 - \frac{j-1}{N}\right)^2 \cdots \left(1 - \frac{1}{N}\right)^2}. \tag{21}$$

If we extend each product in the denominator all the way up to c, and replace the 1 by the product in the numerator, we increase the formula to exactly the Erlang loss function $E(c, a)$. Thus,

$$bl < E(c, a). \tag{22}$$

A similar argument shows that the traffic carried, m, satisfies

$$m < a[1 - E(c, a)]. \tag{23}$$

Thus, for central bus networks the load and loss are bounded above by tbe corresponding Erlang load and loss for c trunks and incoming traffic a.

XX. LINEAR GROWTH IS POSSIBLE IN THE WEAK LIMIT

The central bus concept also leads to an asymptotic estimate of wbat is possible in the way of network growth. We prove

Theorem: For every $\epsilon > 0$, there is an integer c, and a sequence ν_N of networks on N terminals with c central buses such that as $N \to \infty$ with $a \equiv$ constant

$$\begin{aligned}
&(i)\ \ bl(\nu_N) < E(c, a) < \epsilon,\ bl(\nu_N) \to E(c, a)\\
&(ii)\ \ m(\nu_N) < a[1 - E(c, a)],\ m(\nu_N) \to a[1 - E(c, a)]\\
&(iii)\ \ X(\nu_N) \le 136N \log_2 c + O(c)\\
&(iv)\ \ s(\nu_N) \le 4(1 + \log_2 c).
\end{aligned} \tag{24}$$

Remarks: This result says that there exist arbitrarily large networks with specified blocking whose growtb in crosspoints is linear (with slope dependent on blocking), whose complexity in number of stages is logarithmic, and whose load approacbes a constant. Here the growth in "size" N is accompanied by a diminution in the offered load λ per idle pair, according to $\lambda N^2 \equiv a$, a constant, in a natural passage from finite to infinite sources. Because of tbe way these networks will be defined, they will be at the "combinatorially efficient, hard to control" end of the trade-off spectrum.

Proof of Theorem (24): Given $\epsilon > 0$, choose c to be the smallest integer such that $E(c, a) < \epsilon$, and construct a sequence of networks ν_N witb N terminals on a side and c central buses, with nonblocking access to an idle bus from each side. Property (i) follows from (21), (22); (ii) is a result of (23) and the general Erlang formula (1); we have

$$N = \frac{m(1 + \sqrt{1 - \rho})}{1 - \dfrac{m}{a(1 - bl)}}, \tag{4}$$

where m is the load, $a = \lambda N^2$, $bl =$ blocking, and $\rho \in (0, 1)$. Since $m \le c$ we must have, by (4)

$$\frac{n}{a(1 - bl)} \to 1 \quad \text{as} \quad N \uparrow \infty.$$

But by (i), $bl \to E(c, a)$, whence the limit in (ii).

To complete the proof we use the basic bounds on the complexity of nonblocking networks given[5] by Bassalygo and Pinsker, according to wbom each of the $c \times c$ nonblocking networks needed in Fig. 2 can be

made using at most $68c \log_2 c + O(c)$ crosspoints and $2(1 + \log_2 c)$ stages.

XXI. A DIRECT ARGUMENT

It is a property of the "direction" picked for the asymptotics in Theorem (24) that the carried loads m approach a limit, so that the line usage $p = m/N$ goes to zero. In practical terms this means that a fixed amount of traffic is being spread over more and more customers, while the load contribution from any one vanishes. Ultimately, then, this limit "direction" is suitable only for very many lightly loaded lines, and it would be more interesting to have similar or analogous results in which the carried load would increase as the networks grew, and the usage p would be bounded away from zero, with the blocking always less than some prescribed number $\epsilon > 0$.

We know from Theorems (15) and (20) that even without the constraint "$bl < \epsilon$," such a positivity constraint on p necessitates $N \log N$ growth for practical networks. It is instructive, however, to give a separate direct argument for this behavior in the case of the networks constructed in Theorem (24).

Lemma: For $1 > \epsilon > 0$, *let* c *be the earliest integer such that* $E(c, a) \leq \epsilon$. *Then*

$$c \geq a(1 - \epsilon). \tag{25}$$

Proof: By hypothesis, $E(c, a) \leq \epsilon < E(c - 1, a)$. As is well-known, the Erlang function satisfies the recurrence

$$E(c, a) = \frac{1}{1 + \dfrac{c}{aE(c - 1, a)}}.$$

Thus,

$$\frac{1}{1 + \dfrac{c}{aE(c - 1, a)}} \leq \epsilon < E(c - 1, a).$$

Replacing $E(c - 1, a)$ on the left by the smaller ϵ will decrease the left, giving

$$\frac{1}{1 + \dfrac{c}{a\epsilon}} \leq \epsilon.$$

Proposition: Let ν be a central bus network constructed to have $bl \leq \epsilon$, as in Theorem (24), by the method[5] by Bassalygo and Pinsker. Then

$$X(\nu) \geq 8pN \log_2 N + 4Np[\log_2(1 - \epsilon) + \log_2 \lambda]. \tag{26}$$

Proof: ν achieves usage $p = m/N$, so there must be a state $x \in S$ with $|x| \geq Np$. The method of construction implies that ν is a network having $4 + 4 \log_2 c$ stages, where c is the number of central buses, chosen to be the earliest integer c such that $E(c, a) \leq \epsilon$. Thus every call passes through $4 + 4 \log_2 c$ crosspoints. Since two calls do not pass through the same crosspoints, the state x has at least $4Np \log_2 c$ *busy* crosspoints, and so

$$X(\nu) \geq 4Np \log_2 c.$$

But by Lemma (25) and the choice of c, we have

$$\begin{aligned} c &\geq a(1 - \epsilon) = \lambda N^2(1 - \epsilon) \\ \log_2 c &\geq 2 \log_2 N + \log_2(1 - \epsilon) + \log_2 \lambda \qquad (27) \\ X(\nu) &\geq 8pN \log_2 N + 4Np[\log_2(1 - \epsilon) + \log_2 \lambda]. \end{aligned}$$

It follows that any sequence of such networks, growing so that p is bounded away from zero, will grow like N log N. For then λN is bounded below, and (27) implies

$$X(\nu) \geq 2pN \log_2 N + O(N), \quad \text{as} \quad N \to \infty.$$

XXII. FRAME CONCEPT

The second kind of network structure we shall study is called a *frame*. It is familiar to engineers from the No. 5 and earlier crossbar systems. The idea of the frame is to mount all N terminals on a side in k groups of N/k on subnetworks that are connected pairwise by dedicated junctor groups. In the two-sided case shown in Fig. 3, these connections are described by a complete bipartite graph. We shall suppose that the inlet subnetworks are N/k by kc and identical, and are mirror images of the outlet subnetworks, with c trunks or junctors connecting each pair of inlet-outlet subnetworks. A solvable limiting case results when we make the subnetwork nonblocking, so that loss is due always to overload of one of the groups of c junctors between a pair of subnetworks.

Remark: It is tempting to conjecture here that in the natural "weak" limit $N \to \infty$, $\lambda \to 0$, $\lambda N^2 \equiv a$, a constant, the loss for the frame network with nonblocking concentrators will approach the Erlang loss $E(c, ak^{-2})$ for c trunks and Poisson traffic ak^{-2}.

It is not hard to see that if the subnetworks in Fig. 3 are nonblocking, then to define the transition rates that devolve from the stochastic model it is enough to know how many trunks are busy in each of the k^2 dedicated groups of size c. Therefore we can use, instead of our usual microscopic semilattice, a very much reduced[3] notion of state, as follows: we describe the system by a $k \times k$ matrix x of integers (x_{ij}),

$0 \le i, j \le k$, with the interpretation that

x_{ij} = number of calls in progress from subnetwork i on the left to subnetwork j on the right.

The x_{ij} are restricted to be integers between 0 and c, the capacity of each (i, j) trunk group. It is useful to have the notations

$$\begin{aligned} x_i &= \sum_{j=1}^{k} x_{ij} = \text{number of calls in progress on subnetwork } i \text{ on left} \\ x^j &= \sum_{i=1}^{k} x_{ij} = \text{number of calls in progress on subnetwork } j \text{ on right} \\ x^+(i,j) &= \text{the state resulting from } x \text{ when a new call is added to the } (i,j) \text{ trunk group} \\ x^-(i,j) &= \text{the state resulting from } x \text{ when a hangup occurs on the } (i,j) \text{ trunk group.} \end{aligned} \tag{28}$$

With p_x the stationary probability of x, the statistical equilibrium equations are

$$p_x\left[\sum_{i,j=1}^{k} x_{ij} + \lambda \sum_{i,j=1}^{k} 1_{x_{ij}<c}(N - x_i)(N - x^j)\right]$$

$$= \sum_{i,j=1}^{k} [p_{x^+(i,j)}(x_{ij}+1)1_{x_{ij}<c} + \lambda 1_{x_{ij}>0} p_{x^-(i,j)}(N - x_i + 1)(N - x^j + 1)].$$

Here the indicator functions for $x_{ij} \neq 0$ and $x_{ij} \neq c$ give the right equations on the "boundary" of the state space. The solution of these equations has the convenient product form

$$\begin{aligned} p_x &= p_0 \frac{\lambda^{|x|}}{\prod_{i,j} x_{ij}!} \prod_{i,j} x_i! x^j! \binom{N}{x_i}\binom{N}{x^j} \\ &= p_0 \lambda^{|x|} \prod_{i,j} \binom{x_i}{x_{i1}, \cdots, x_{ik}}\binom{x^j}{x_{1j}, \cdots, x_{kj}}\binom{N}{x_i}\binom{N}{x^j}, \end{aligned}$$

where p_0 is the chance that no calls are up, given as the reciprocal of the normalizing sum as

$$p_0^{-1} = \sum_x \lambda^{|x|} \prod_{i,j} \binom{x_i}{x_{i1}, \cdots, x_{ik}}\binom{x^j}{x_{1j}, \cdots, x_{kj}}\binom{N}{x_i}\binom{N}{x^j}.$$

The sum over the states x is over all $k \times k$ matrices whose entries are integers in [0, c], including the identically 0 state, which contributes a term 1.

Under our symmetry assumptions the probability of blocking between two outer subnetworks i and j is the same as the overall loss,

and is given by

$$bl = \frac{\sum_x p_x 1_{x_{ij}=c}(N - x_i)(N - x^j)}{\sum_x p_x(N - x_i(N - x^j)}.$$

This formula does not depend on the normalizer p_0^{-1}. Although it is exact, it will be of interest to us primarily for the insight it gives into the asymptotic behavior of bl as $N \to \infty$, $\lambda \to 0$, with $\lambda N^2 = a$, constant. To this end it is enough to look at p_x/p_0. Writing it in the form, for $x \neq 0$, and using $\sum_i x_i = \sum_j x^j = |x|$,

$$p_x/p_0 = \frac{\lambda^{|x|}}{\prod_{i,j} x_{ij}} \prod_{i,j} \prod_{m=0}^{x_i - 1} \left(\frac{N}{k} - m\right) \prod_{\ell=0}^{x^j - 1} \left(\frac{N}{k} - \ell\right)$$

$$= \left(\frac{\lambda N^2}{k^2}\right)^{|x|} \cdot \prod_{i,j} \cdot \frac{\left(1 - \frac{x_i - 1}{N/k}\right) \cdots \left(1 - \frac{1}{N/k}\right)\left(1 - \frac{x^j - 1}{N/k}\right) \cdots \left(1 - \frac{1}{n/K}\right)}{x_{ij}!}$$

(the products interpreted as 1 if x_i or x^j is 0) one can see that

$$\lim_{\substack{N \to \infty \\ \lambda \to 0 \\ \lambda N^2 = a}} p_x/p_0 = \frac{(ak^{-2})^{|x|}}{\prod_{i,j} x_{ij}!};$$

write $\alpha = ak^{-2}$ for simplicity; the probability of loss goes to

$$\frac{\frac{\alpha^c}{c!} \sum_{x_{ij}=c} \frac{\alpha^{|x|-c}}{\prod_{k,\ell \neq i,j} x_{k\ell}!}}{\sum_{n=0}^{c} \frac{\alpha^n}{n!} \sum_{x_{ij}=n} \alpha^{|x|-n} \Big/ \prod_{k,\ell \neq i,j} x_{\ell k}!}.$$

It can be verified that for the various n up to c, the $\sum_{x_{ij}=n}$ summands in the denominator are all equal, and equal to the $\sum_{x_{ij}=c}$ sum in the numerator. Thus, as conjectured,

$$\lim_{\substack{n \to \infty \\ \lambda \to 0 \\ \frac{\lambda N^2}{k^2} = a}} bl = \frac{\frac{\alpha^c}{c!}}{\sum_{n=0}^{c} \frac{\alpha^n}{n!}} = E(c, \alpha) = E(c, ak^{-2}). \tag{29}$$

Table I—Load loss relationships

	Loss	Load
Central bus	$E(c, a)$	$a[1 - E(c, a)]$
Frame	$E(c, ak^{-2})$	$a[1 - E(c, ak^{-2})]$

As for central bus networks, one can get linear growth in the weak limit, described as follows:

Theorem: For every $\epsilon > 0$, there is an integer c and a growing sequence ν_N of two-sided frame networks on N terminals, with N/k subnetworks on a side, such that as $N \to \infty$, $\lambda \to 0$, and $a = \lambda N^2 \equiv$ constant,

$$\begin{aligned} &(i)\ bl(\nu_N) \to E(c, ak^{-2}) \leq \epsilon \\ &(ii)\ m(\nu_N) \to a[1 - E(c, ak^{-2})] \\ &(iii)\ X(\nu_N) \leq 2k[68N \log_2 c + O(c)] \\ &(iv)\ s(\nu_N) = 4(1 + \log_2 kc). \end{aligned} \tag{30}$$

Remark: The reader can check that it is quite proper to regard the frame network with k concentrating subnetworks on a side as a system of k^2 central bus networks: the loss is asymptotically Erlang $E(c, \cdot)$ in both cases, and the carried load for the frame is k^2 times the carried load on the corresponding central bus, as shown in Table I.

Proof of Theorem (30): This proof is very much like that of (24). We chose c to be the least integer such that $E(c, ak^{-2}) \leq \epsilon$, and construct nonblocking $N/k \times kc$ concentrators for the subnetworks, each using no more than $68(N/k)\log_2(kc) + O(kc)$ crosspoints and $2(1 + \log_2 kc)$ stages. Convergence of the loss has been proved as (29), and that of the load follows as usual from the Erlang formula (1), as in (24).

XXIII. REMOTE CONCENTRATOR CONCEPT

A third network structure worth looking at consists of concentrating subnetworks each connected only to one and the same central "core" network by a group of c high-usage trunks, as in Fig. 4. We call this structure, well-known to traffic engineers, the "remote concentrator concept;" it represents an extreme form of the advice to separate concentration and distribution in the network. We shall suppose that the subnetworks are divided into two groups, one carrying the inlets, the other the outlets, so that the whole network is still two-sided. When the outer subnetworks (concentrators) and the inner core are all nonblocking, a useful solvable case results, and we can again guess that as the right limit is taken there will be some connection with Erlang's E function. Since the success of a call attempt depends

entirely on finding first a free trunk into the core, and then one going out from it to the destination concentrator, it is tempting to guess that asymptotically the loss is given hy the "blocking polynomial" $1 - (1 - b)^2$, where b is the chance that all c trunks in a group are husy. This simple guess is prohahly not true, because of the correlation between loads on trunk groups; it may nevertheless be a hound, although we have not heen able to prove this: the question wbether the blocking owing to simultaneously full groups is larger or smaller than b^2 is open. We can, however, give Erlang E hounds on hoth kinds of hlocking, as well as exact loss formulas for finite N in terms of logarithmic derivatives of a partition function.

Let k he a divisor of N, fixed henceforth, to he interpreted as the number of concentrators on the inlet side of the network, each with N/k inlets, c trunks to the core, and nonblocking. The outlet side is similarly constituted. As a notion of state we can take the matrices $x = (x_{ij}, 0 \le i, j \le k)$ with the meaning

x_{ij} = numher of calls in progress
from inlet concentrator i to outlet concentrator j.

These matrices are subject to the condition that hoth rows and columns must not sum to more than c, the trunk group size. This is the same notion of state used for frame networks, except that there the *entries* bad to he at the most c, while here the row and column sums are tbus hounded.

Using the same notations (28) as for the frame networks, we can put down the following equilibrium equations:

$$p_x \left[\sum_{i,j=1}^{k} x_{ij} + \lambda \sum_{i,j=1}^{k} 1_{x_i<c, x^j<c} \left(\frac{N}{k} - x^j \right) \right]$$
$$= \sum_{i,j=1}^{k} \left[p_{x^+(i,j)}(x_{ij} + 1) 1_{x_{ij}<c} + \lambda 1_{x_{ij}>0} p_{x^-(i,j)} \left(\frac{N}{k} x_{i+1} \right) \left(\frac{N}{k} - x^j + 1 \right) \right].$$

Again, the indicator factors give the right equations at the houndary of the state space, and the solution bas the same product form as for frame networks:

$$p_x = p_0 \lambda^{|x|} \prod_{i,j=1}^{k} \binom{x}{x_{i1}, \cdots, x_{ik}} \binom{x^j}{x_{ij}, \cdots, x_{kj}} \binom{N/k}{x_i} \binom{N/k}{x^j},$$

where p_0 is the cbance of no calls in progress, the reciprocal of the normalizer

$$\sum_{x \in S} \lambda^{|x|} \prod_{i,j=1}^{k} \binom{x_i}{x_{i1}, \cdots, x_{ik}} \binom{x^j}{x_{ij}, \cdots, x_{kj}} \binom{N/k}{x_i} \binom{N/k}{x^j}.$$

The sum over tbe states $x \in S$ is over all $k \times k$ matrices with nonnegative integer entries, and row *and* column sums at most c.

By symmetry, the probability of blocking between two remote concentrators i and j is the same as the overall loss, and is given by

$$bl = \frac{\sum_{x \in S} p_x [1_{x_i = c} \wedge 1_{x^j = c}] \left(\frac{N}{k} - x_i\right)\left(\frac{N}{k} - x^j\right)}{\sum_{x \in S} p_x \left(\frac{N}{k} - x_i\right)\left(\frac{N}{k} - x^j\right)}.$$

This formula depends only on the known ratios p_x/p_0. We examine its asymptotic comportment as $\lambda \to 0$, $N \to \infty$, with $\lambda N^2 \equiv a$, constant. The argument for (29) gives

$$\lim_{\substack{N \to \infty \\ \lambda \to 0 \\ \lambda N^2 = a}} p_x/p_0 = \frac{\left(\frac{a}{k^2}\right)^{|x|}}{\prod_{i,j=1}^{k} x_{ij}!}.$$

XXIV. THE PARTITION FUNCTION

We have, for the remote concentrator concept,

$$p_x = p_0 \lambda^{|x|} \prod_{i,j=1}^{|x|} \binom{x_i}{x_{i1}, \cdots, x_{ik}} \binom{x^j}{x_{1j}, \cdots, x_{kj}} \binom{N/k}{x_i} \binom{N/k}{x^j}$$

$$= p_0 \lambda^{|x|} c(x).$$

Hence, introducing the generating function

$$\phi(y) = 1 + \sum_{j=1}^{kc} y^j \sum_{\substack{|x| = j \\ x \in S}} c(x),$$

the moments of the number of calls in progress can be expressed as logarithmic derivatives; especially,

$$m = \lambda \frac{d}{d\lambda} \log \phi(\lambda)$$

$$\sigma^2 = \lambda^2 \frac{d^2}{d\lambda^2} \log \phi(\lambda) + \lambda^2 \frac{d}{d\lambda} \log \phi(\lambda),$$

and by the generalized Erlang formula (1), the blocking is determined via

$$1 - bl = \frac{1}{\lambda} \frac{m}{(N - m)^2 + \sigma^2}$$

$$= \frac{\frac{d}{d\lambda} \log \phi(\lambda)}{\left[N - \lambda \frac{d}{d\lambda} \log \phi(\lambda)\right]^2 + \lambda^2 \frac{d^2}{d\lambda^2} \log \phi(\lambda) + \lambda^2 \frac{d}{d\lambda} \log \phi(\lambda)}.$$

The interested reader can verify that analogous results hold in the Erlang case of c trunks offered traffic a: the generating function is

$$\phi(y) = 1 + y + \frac{y^2}{2} + \cdots + \frac{y^c}{c!}$$

so that

$$1 - bl = 1 - E(c, a) = \frac{\phi'(a)}{\phi(a)}$$

$$m = a(1 - E(c, a)) = a \frac{d}{da} \log \phi(a).$$

The essential form of these relationships persists in the weak limit $\lambda \to 0$, $N \to \infty$, $\lambda N^2 = a \equiv$ constant. We write

$$p_x = p_0 a^{|x|} \prod_{i,j=1}^{k} \binom{x_i}{x_{i1}, \cdots, x_{ik}} \binom{x^j}{x_{1j}, \cdots, x_{kj}} \frac{(N/k)!(N/k)!N^{-2|x|}}{\left(\frac{N}{k} - x_i\right)!x_i!\left(\frac{N}{k} - x^j\right)!x^j!}$$

and take the limit as above. Stirling's formula implies that the partition function becomes

$$\phi(y) = 1 + \sum_{\ell=1}^{kc} y^\ell \sum_{\substack{|x|=\ell \\ \text{row sums} \le c \\ \text{column sums} \le c}} k^{-2\ell} \cdot \prod_{i,j=1}^{k} \frac{\binom{x_i}{x_{i1}, \cdots, x_{ij}} \binom{x^j}{x_{1j}, \cdots, x_{kj}}}{x_i! x^j!}. \quad (31)$$

Then since $(N - m)^2 + \sigma^2 \sim N^2$ in the weak limit, one finds

$$m \to a \frac{d}{da} \log \phi(a) \quad (32)$$

$$1 - bl \to \frac{d}{da} \log \phi(a). \quad (33)$$

Theorem: For every integer k and every $\epsilon > 0$, there is an integer c, and a sequence ν_N of "remote concentrator" networks, with c trunks from each of k $N/k \times c$ concentrators on a side to a central core $kc \times kc$, such that as $\lambda \to 0$, $N \to \infty$, $\lambda N^2 = a \equiv$ constant

$$\begin{aligned}
&(i)\ bl(\nu_N) \to 1 - \frac{d}{da} \log \phi(a) < \epsilon \\
&(ii)\ m(\nu_N) > a \frac{d}{da} \log \phi(a) \\
&(iii)\ X(\nu_N) \le \frac{2N}{k} [68c \log_2 c + O(c)] + 68kc \log_2 kc + O(kc) \\
&(iv)\ s(\nu_N) = 6(1 + \log_2 c) + 2 \log_2 k.
\end{aligned} \quad (34)$$

Proof: Again, this proof is like that of (24). Using (32), the blocking can be written as

$$bl = 1 - \frac{m}{\lambda N^2 \sum_{x \in S} p_x \left(1 - \frac{x_i}{N}\right)\left(1 - \frac{x^j}{N}\right)} = 1 - \frac{m}{a + o(1)}$$

$$= 1 - \frac{d}{da} \log \phi(a) + o(1).$$

So with $\lambda N^2 = a$, fixed, pick the integer c in the partition function $\phi(y)$ defined by (31) so that

$$1 - \frac{d}{da} \log \phi(a) < \epsilon.$$

This is possible because the limit blocking must decrease to 0 with increasing c. Thus, (32) proves (*i*) and (*ii*); (*iii*) and (*iv*) follow by the same kind of concentrator construction as before, using the method of Bassalygo and Pinsker.[5]

XXV. BLOCKING INEQUALITIES FOR REMOTE CONCENTRATOR CONCEPT

When the concentrating subnetworks and the core are nonblocking, it is possible to derive some interesting Erlang E bounds for the blocking in a remote concentrator structure. We first note tbat the contributions to blocking are of two kinds: a call is blocked between subnetworks i and j because $x_i = c$, or $x^j = c$, or both, i.e., $x_i + x^j = 2c$. Thus,

$$bl = \frac{\sum_{x \in S} p_x \left(\frac{N}{k} - x_i\right)\left(\frac{N}{k} - x^j\right)(1_{x_i=c} + 1_{x^j=c} - 1_{x_i+x^j=2c})}{\sum_{x \in S} p_x \left(\frac{N}{k} - x_i\right)\left(\frac{N}{k} - x^j\right)}$$

$$= \sum_{x \in S} p_x (1_{x_i=c} + 1_{x^j=c} - 1_{x_i+x^j=2c}) + o(1).$$

By symmetry the $1_{x_i=c}$ and $1_{x^j=c}$ terms contribute the same amount, so the problem of estimating tbe blocking reduces to estimating:

(*i*) the probability $Pr\{x_i = c\} = \sum_{x_i=c} p_x$ of having "all trunks busy" on concentrator i, and

(*ii*) the "double trouble" term $Pr\{x_i + x^j = 2c\}$.

It is convenient to define, for states $x \in S$, and $1 \le i, j \le k$,

$$s^{ij}(x) = (1 - 1_{x_i=c})(1 - 1_{x^j=c})\left(\frac{N}{K} - x_i\right)\left(\frac{N}{k} - x^j\right).$$

This is the number of unblocked idle inlet-outlet pairs (u, v) with u on concentrator i and v on concentrator j.

Lemma: For integers $0 \le t \le w = max_{x \in S}|x|$, the chance of t calls up on concentrator i can be represented as

$$Pr\{x_i = t\} = Pr\{x_i = 0\} \frac{a^t}{t!} \prod_{\ell=0}^{t-1} \eta_\ell, \tag{35}$$

where

$$\eta_\ell = N^{-2\ell} \sum_{x_i=\ell} \frac{p_x}{Pr\{x_i = \ell\}} \sum_{j=1}^{k} s^{ij}(x)$$

$$= N^{-2\ell} E\{\text{number of unblocked calls on } i | \ell i\text{-trunks are busy}\}$$

$$\le k^{-1}.$$

Proof: This follows from the form of the statistical equilibrium equations, which says that the rate into a set is the rate out of it. We use the sets $\{x \in S: x_i = t\}$, which partition S and communicate by pairs in a simply ordered array except at the endpoints $\ell = 0$ and $\ell = c$. The result follows by iteration. In a similar way it is found that

Lemma:

$$Pr\{x_i + x^j = t\} = Pr\{x_i + x^j = 0\} \frac{a^t}{t!} \prod_{\ell=0}^{t-1} \xi_\ell, \tag{36}$$

where

$$\xi_\ell = N^{-2\ell} \sum_{x_i + x^j = \ell} \frac{p_x}{Pr\{x_i + x^j = \ell\}} \sum_{m=1}^{k} [s^{im}(x) + (1 - \delta_{im}) s^{mj}(x)]$$

$$N^{-2\ell} E\{\text{number of unblocked calls to } i \text{ or } i | x_i + x^j = \ell\}$$

$$\le \frac{2k - 1}{k^2}.$$

Theorem:

$$Pr\{x_i = c\} \le E(c, ak^{-2})$$

$$Pr\{x_i + x^j = 2c\} = E(2c, 2ak^{-1} - ak^{-2}). \tag{37}$$

Proof: Introduce the function on the positive orthant

$$f(y_1, \cdots, y_c) = \frac{y_1 y_2 \cdots y_c}{1 + y_1 + y_1 y_2 + \cdots + y_1 y_2 \cdots y_c} = \frac{N}{D};$$

this is increasing in each y_i there, since

$$\frac{\partial f}{\partial y_i} = \frac{N(1 + y_2 + y_2 y_2 + \cdots + y_1 y_2 \cdots y_{i-1})}{y_i D^2} > 0.$$

By normalization, $\sum_{i=0}^{c} Pr\{x_i = t\} = 1$, so Lemma (35) says that

$$Pr\{x_i = c\} = f\left(a\eta_0, \frac{a}{2}\eta_1, \cdots, \frac{a}{c}\eta_{c-1}\right)$$

$$\leq f\left(a/k, \frac{a/k}{2}, \cdots, \frac{a/k}{c}\right) = E(c, a/k).$$

The proof for the "double trouble" term is analogous, from Lemma (36).

XXVI. CONCLUSIONS AND PROSPECTS

For the narrow class of probability models for telephone networks described by "finite sources, exponential holding times," we have shown that the $N \log N$ rate of growth (of the number X of crosspoints) characteristic of nonblocking networks extends also to those with blocking. This narrow class provided easy methodological devices for carrying out the proofs. Extensions to more general statistics have been mentioned in an interesting series of papers by N. Pippenger, listed in the bibliography. However, his results are either combinatorial or restricted to Markovian models similar to ours. Since some of his principal demonstrations depend on what amounts to the old "lost calls held" convention applied to finite sources, his results are strictly not comparable to those given here. Extensions to the distribution-free context remain to be made. As Pippenger suggests, the most useful tools are likely to be the entropy concept and ergodic theory.

REFERENCES

1. E. Brockmeyer, H. L. Halström, and A. Jensen, *The life and work of A. K. Erlang*, Acta Polytechnica Scandinavica, Mathematics and Computing Machinery Series, No. 6, Copenhagen, 1960.
2. V. E. Beneš, "Markov Processes Representing Traffic in Connecting Networks," B.S.T.J., *42*, No. 6 (November 1963) pp. 2795–838.
3. V. E. Beneš, "Reduction of Network States Under Symmetries," B.S.T.J., *57*, No. 1 (January 1978), pp. 111–49.
4. C. Clos, "A Study of Nonblocking Switching Networks, B.S.T.J., *32*, No. 2 (March 1953), pp. 406–24.
5. L. A. Bassalygo and M. S. Pinsker, "Complexity of an Optimal Nonblocking Switching Network Without Reconnections," Problemy Peredachi Informatsii, *9* (1973), pp. 84–7; translated into English in Problems of Information Transmission, *9* (1974), pp. 64–6.

BIBLIOGRAPHY OF BACKGROUND READING AND RELATED WORK

Beneš, V. E., *Mathematical Theory of Connecting Networks and Telephone Traffic*, New York: Academic Press, 1965.

Ikeno, N., "A Limit on Crosspoint Number," 1959 International Symposium on Circuit and Information Theory, Los Angeles, CA, June 16–18, 1959.

Pippenger, N., "The Complexity Theory of Switching Networks," Sci. D. Thesis, Mass. Inst. Technology, 1973.

Pippenger, N., "Complexity of Seldom Blocking Networks," Proc. IEEE Communications Conference, 1976, paper 7–8.

Pippenger, N., "On the Complexity of Strictly Nonblocking Networks," IEEE Trans. Communications, *COM-22* (1974), pp. 1890–2.

Shannon, C. E., "Memory Requirements in a Telephone Exchange, B.S.T.J., *29*, No. 3 (July 1950), pp. 343–9.